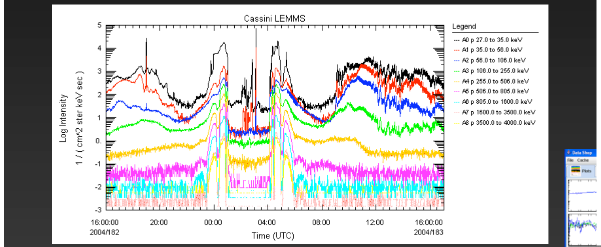
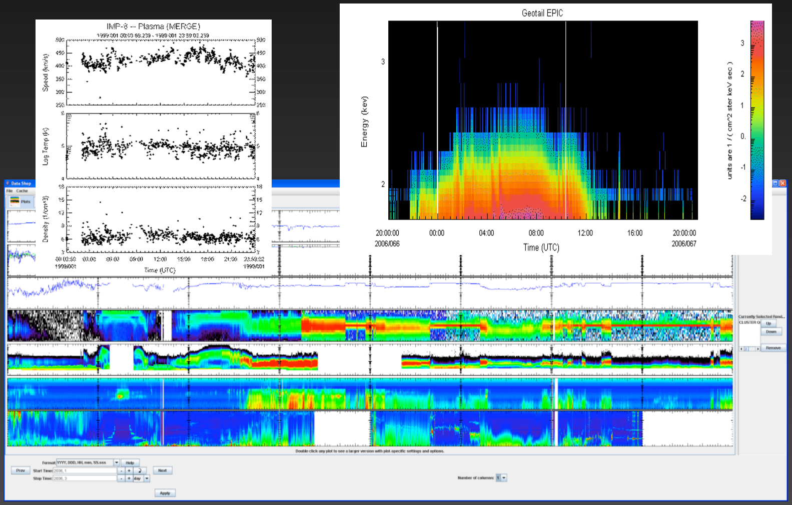
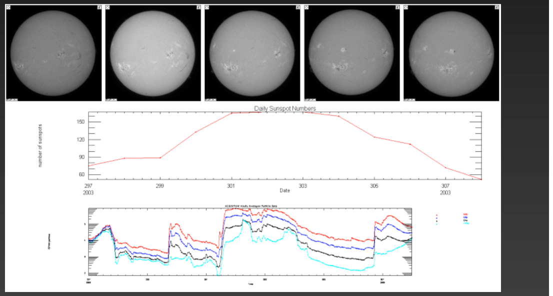
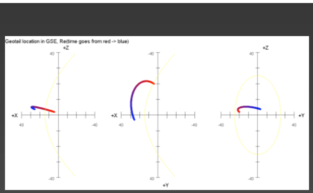
Groundwork for Integrated Analysis of Distributed S3C Data









1. *steps in doing integrated analysis*

2. *how do you integrate so many diverse resources?*
   hint: don't rely too much on meta-data

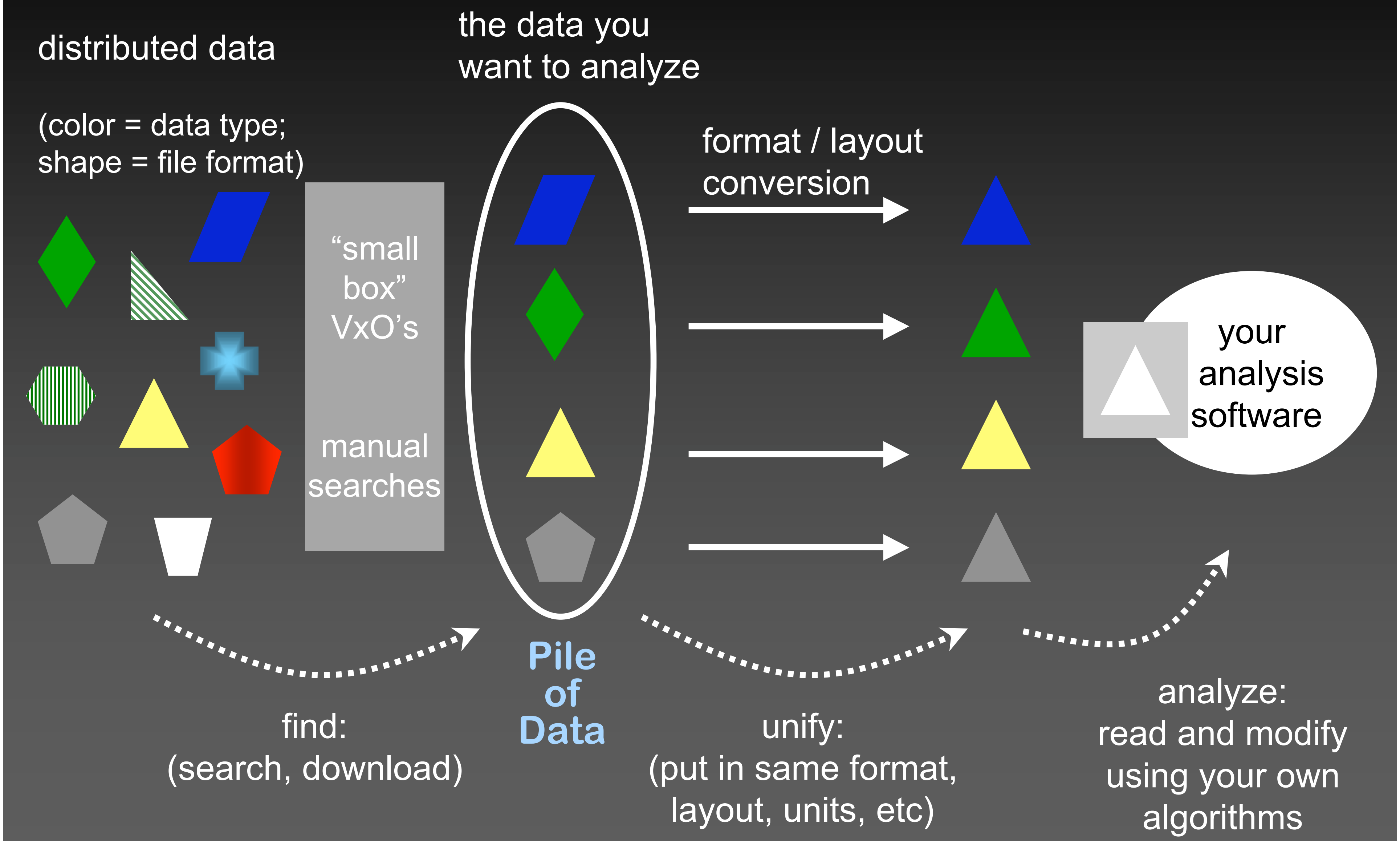
3. *analysis and beyond – what the groundwork will enable*

4. *examples of first generation capabilities*

Jon Vandegriff
JHU / APL
Aaron Roberts
Adam Szabo
Goddard Space Flight Center

SM23A-01
Spring AGU May 23, 2006

Definition: integrated analysis



distributed data

(color = data type;
shape = file format)

“small box” VxO’s

manual searches

the data you want to analyze

Pile of Data

format / layout conversion

your analysis software

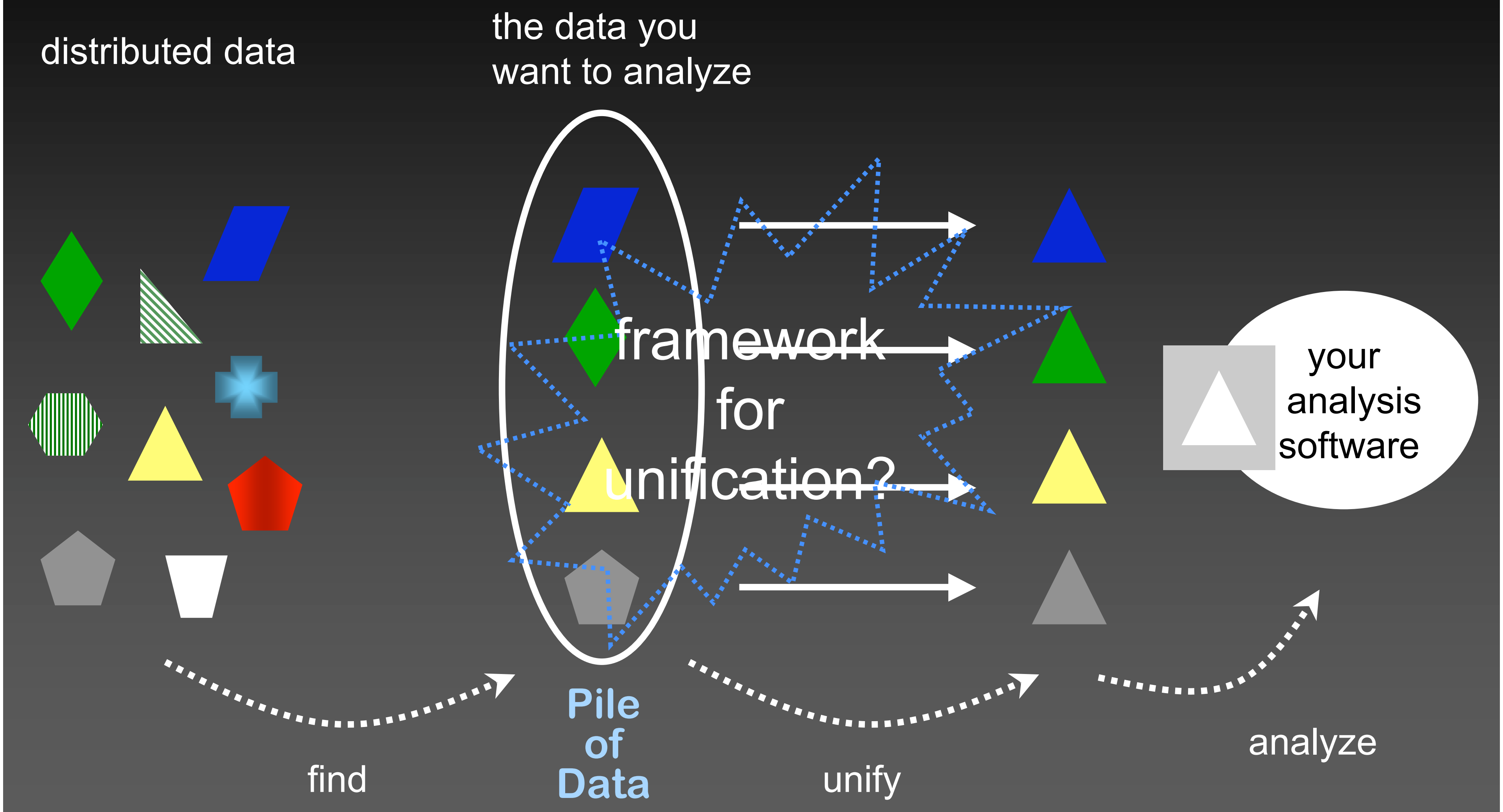
find:
(search, download)

unify:
(put in same format, layout, units, etc)

analyze:
read and modify using your own algorithms

Jon Vandegriff -- JHU / APL

Focus on the Unification Problem



distributed data

the data you
want to analyze

framework
for
unification?

your
analysis
software

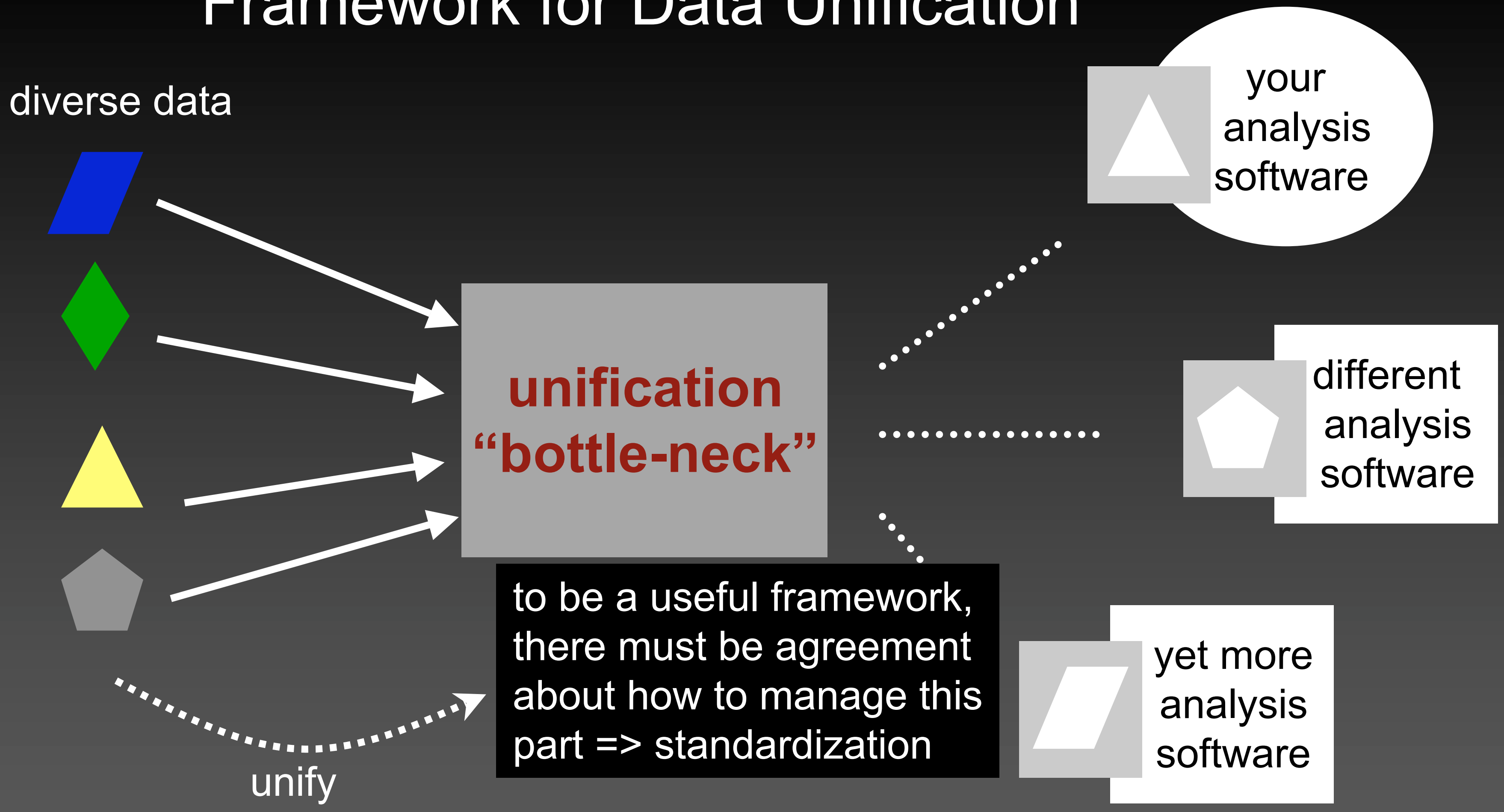
find

Pile
of
Data

unify

analyze

Framework for Data Unification



diverse data

unification “bottle-neck”

to be a useful framework, there must be agreement about how to manage this part => standardization

unify

your analysis software

different analysis software

yet more analysis software

Interoperability among a large number of sources and destinations *requires* a standard data layout somewhere in the chain.
At some point, the data all has to be accessible through exactly the same STANDARDIZED mechanism – such a “unification bottle-neck” is unavoidable.

Minimizing the Pain of Standardization

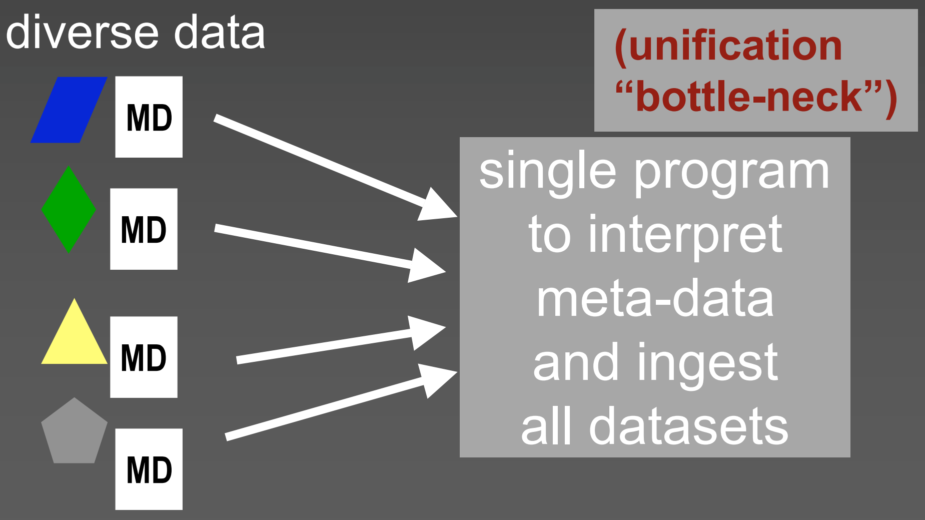
What options exist for the “bottle-neck”:

1. Require a common file format? Very appealing: one set of readers to access everything!

   which format? who does translation? keeping the copies up to date?

2. Use meta-data to describe access details:

   define an XML schema to describe all possible file formats and layouts;
   then one piece of software uses the meta-data to interpret anything


diverse data
MD
MD
MD
MD
(unification “bottle-neck”)
single program to interpret meta-data and ingest all datasets

MD = different meta-data for each dataset

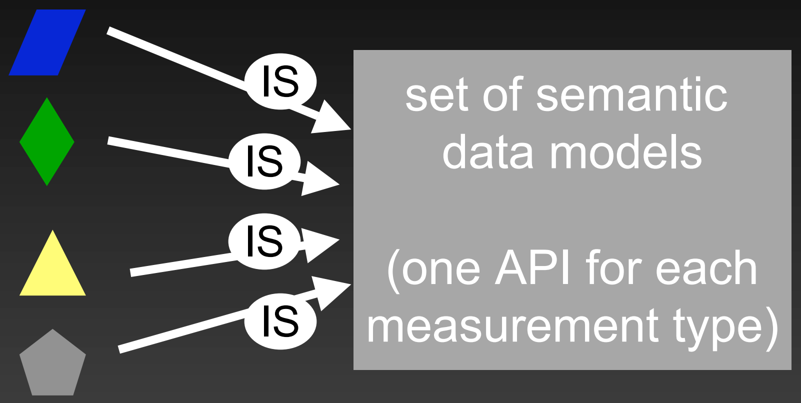
problems:

1. the complexity involved in handling file content is significantly higher than what has been tackled so far with SPASE; long time to come to consensus

2. XML Schema – not complex enough; many datasets will have features which just can’t be captured by an XML standard

Minimizing the Pain of Standardization (continued)

A better option:
“Data Interfaces”



set of semantic data models

(one API for each measurement type)

(unification “bottle-neck”)

IS = interpreter software for each dataset

=> delay the standardization process until the last possible moment!

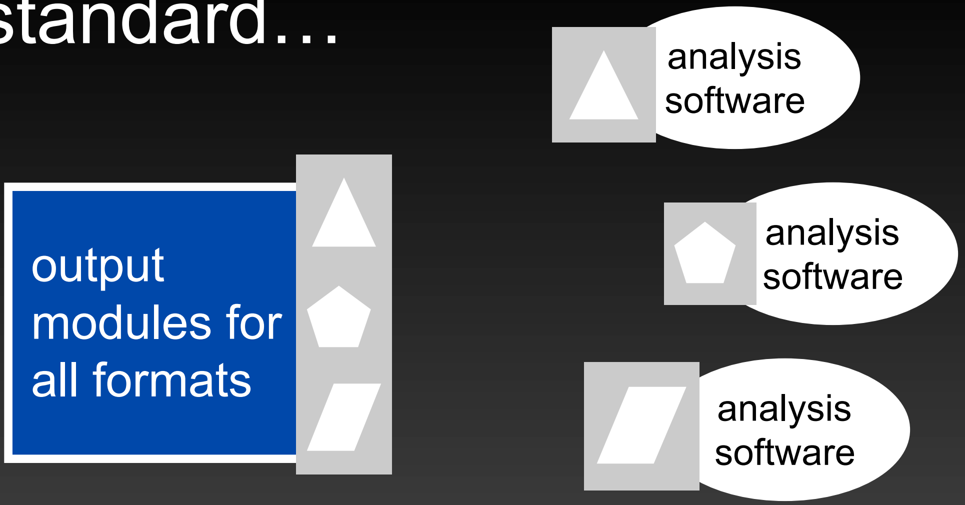
create a standardized model of the content in each type of science measurement: MAG, Plasma, Particles, Waves, S/C location, S/C pointing, etc

for each dataset, write one piece of interpreter software that knows how to make the dataset conform to the appropriate standard interface

its not a standard file format – it’s a standard that’s internal to the software; its an “overlay” of accessor methods that get applied after the data is read

Once you have a data standard…

key point: one set of readers can now access everything (same benefit as with requiring a common file format)

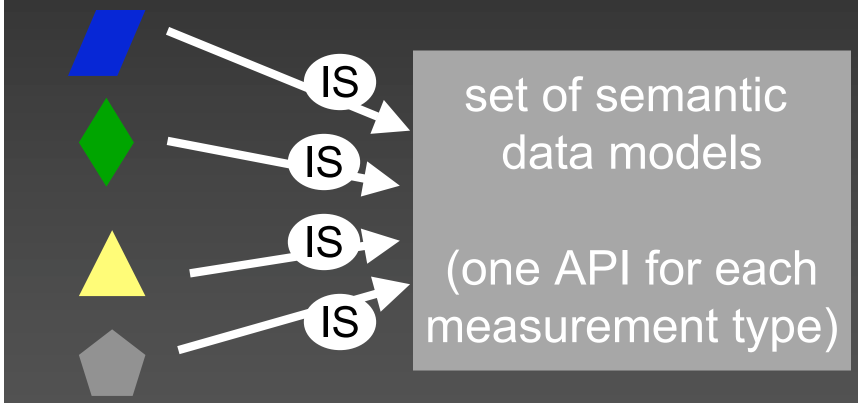


output modules for all formats

analysis software

analysis software

analysis software



set of semantic data models

(one API for each measurement type)

shared software library for visualization and analysis

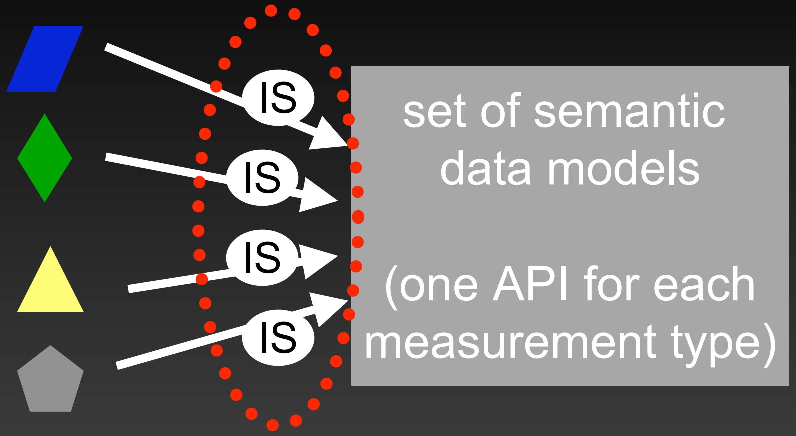
IDL routines similar to SolarSoft

web-based browsing tools to help with data discovery

semantic description of the data can be connected to emerging ontologies

data mining servers are now easier to develop (“huge box” VO)

Writing Interpreter Software


IS
IS
IS
IS
set of semantic data models
(one API for each measurement type)

This framework requires accessor / interpreter software to be written for every dataset to be included.

Estimated time: few hours to 1 week (depending on organization of dataset)

Does not need to be written by the data provider!

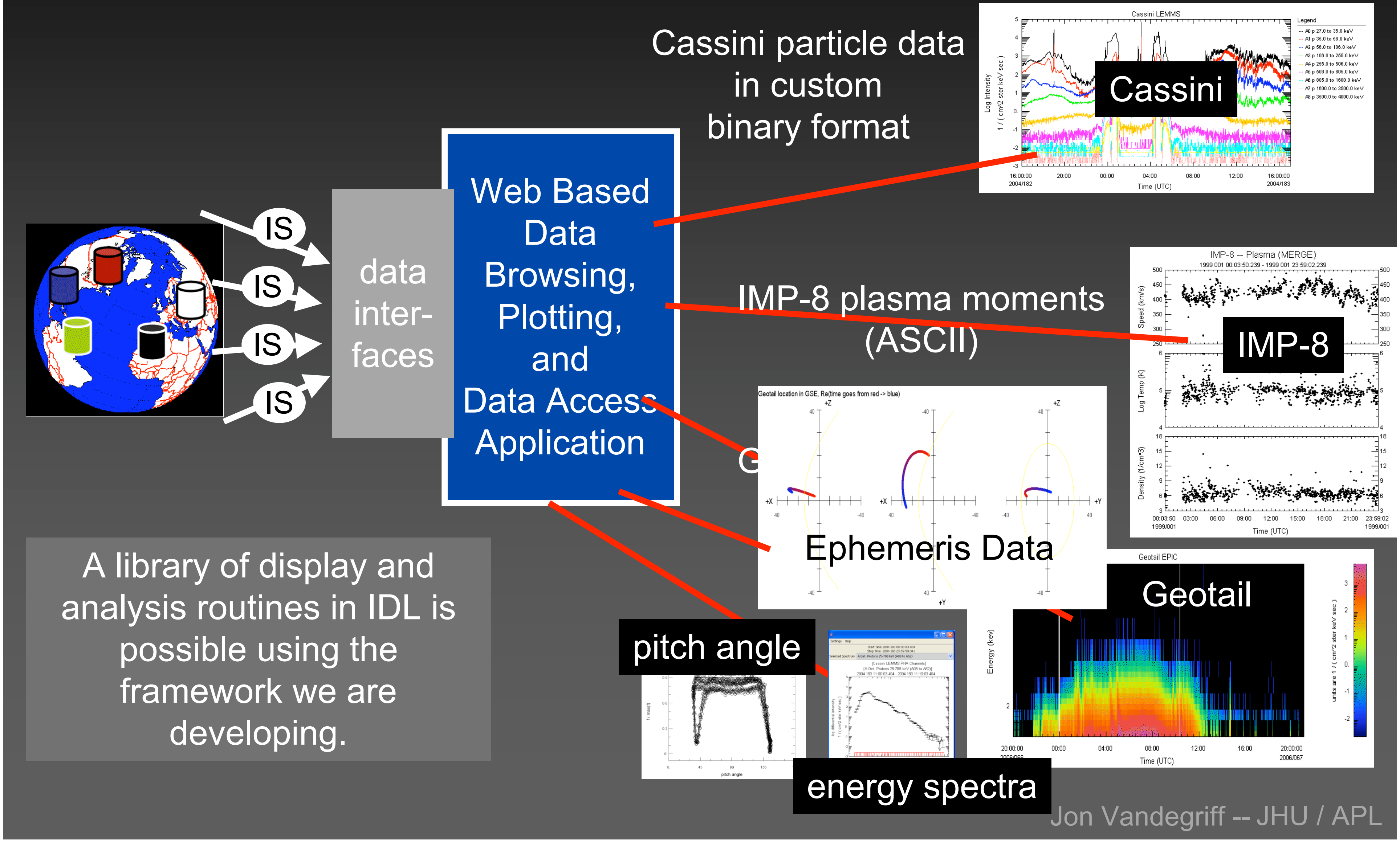
We are preparing a document describing what is needed from providers in order to be able to write these accessors.

We will be developing accessor / interpreter code modules which can be meta-data driven. The meta-data for these will be easier to specify, since its scope is restricted to a limited set of data types.

Proof of Concept - MIDL

http://sd-www.jhuapl.edu/MIDL

Plotting tools and data access for diverse and distributed resources.



IS

IS

IS

IS

data inter-faces

Web Based Data Browsing, Plotting, and Data Access Application

Cassini particle data in custom binary format

Cassini

IMP-8 plasma moments (ASCII)

IMP-8

Ephemeris Data

Geotail

pitch angle

energy spectra

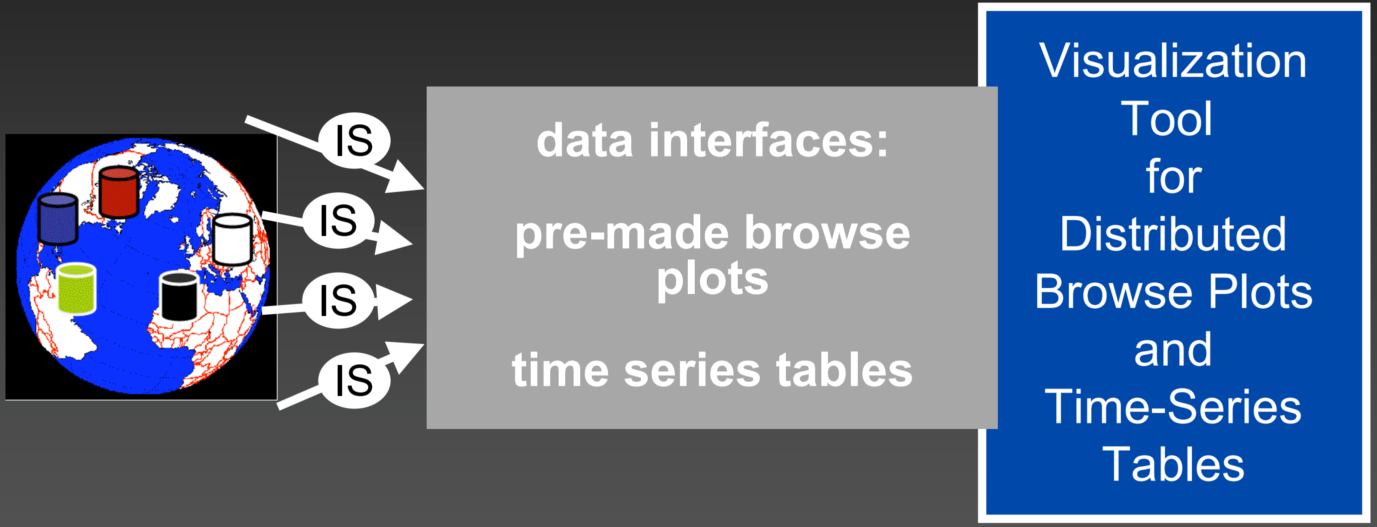
A library of display and analysis routines in IDL is possible using the framework we are developing.

Jon Vandegriff -- JHU / APL

Proof of Concept - DataShop

`http://sd-www.jhuapl.edu/datashop`

Combination browse plot viewer and time-series plotter.



data interfaces:

pre-made browse plots

time series tables

Visualization Tool for Distributed Browse Plots and Time-Series Tables

One place to view browse plots for multiple missions -
a simplistic but useful kind of integration.

Proof of Concept - DataShop

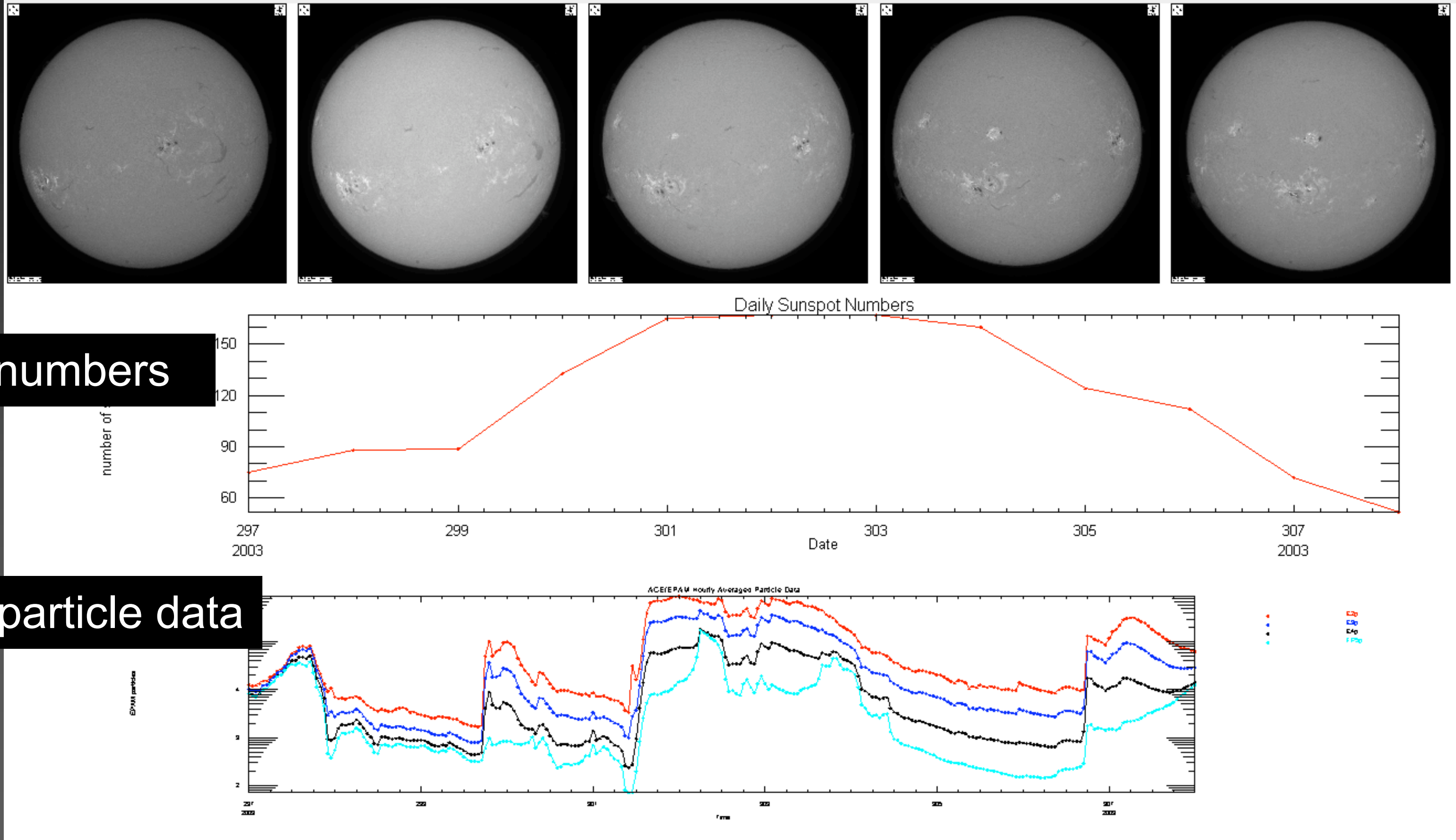
`http://sd-www.jhuapl.edu/datashop`

Integrates distributed resources: pre-made images and time-series data.

solar images

daily sunspot numbers

ACE/EPAM particle data



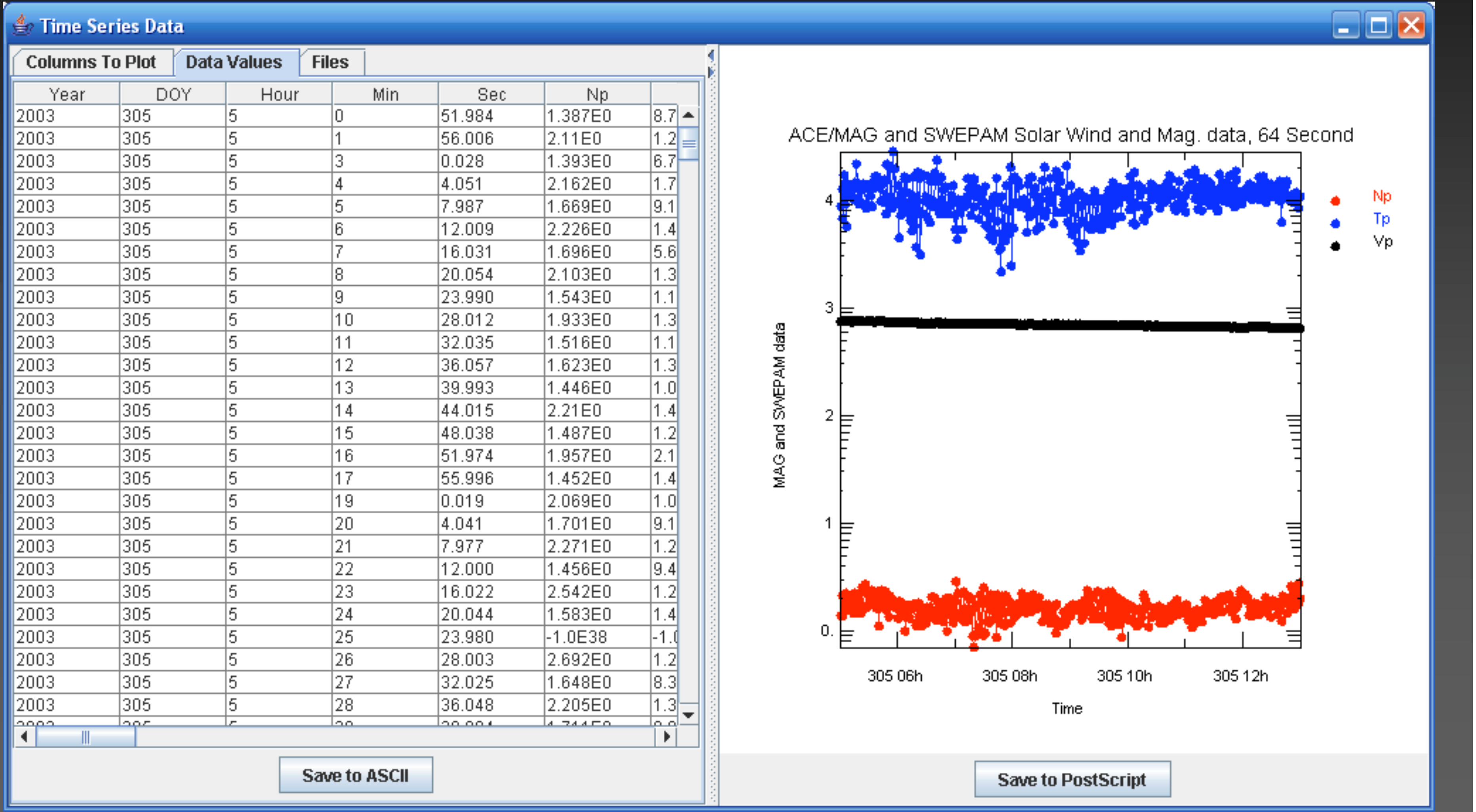
Stacked plot of several diverse data types, all accessed remotely.

Jon Vandegriff -- JHU / APL

Proof of Concept - DataShop

http://sd-www.jhuapl.edu/datashop

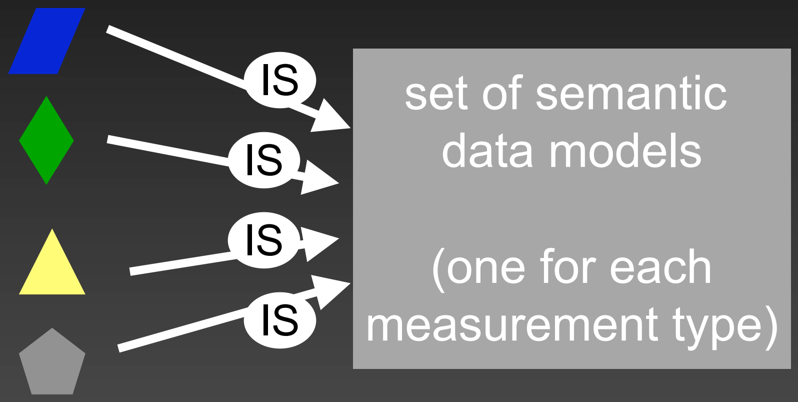
For the time series data, you can customize the plot, and get to the data.



Year	DOY	Hour	Min	Sec	Np	
2003	305	5	0	51.984	1.387E0	8.7
2003	305	5	1	56.006	2.11E0	1.2
2003	305	5	3	0.028	1.393E0	6.7
2003	305	5	4	4.051	2.162E0	1.7
2003	305	5	5	7.987	1.669E0	9.1
2003	305	5	6	12.009	2.226E0	1.4
2003	305	5	7	16.031	1.696E0	5.6
2003	305	5	8	20.054	2.103E0	1.3
2003	305	5	9	23.990	1.543E0	1.1
2003	305	5	10	28.012	1.933E0	1.3
2003	305	5	11	32.035	1.516E0	1.1
2003	305	5	12	36.057	1.623E0	1.3
2003	305	5	13	39.993	1.446E0	1.0
2003	305	5	14	44.015	2.21E0	1.4
2003	305	5	15	48.038	1.487E0	1.2
2003	305	5	16	51.974	1.957E0	2.1
2003	305	5	17	55.996	1.452E0	1.4
2003	305	5	19	0.019	2.069E0	1.0
2003	305	5	20	4.041	1.701E0	9.1
2003	305	5	21	7.977	2.271E0	1.2
2003	305	5	22	12.000	1.456E0	9.4
2003	305	5	23	16.022	2.542E0	1.2
2003	305	5	24	20.044	1.583E0	1.4
2003	305	5	25	23.980	-1.0E38	-1.0
2003	305	5	26	28.003	2.692E0	1.2
2003	305	5	27	32.025	1.648E0	8.3
2003	305	5	28	36.048	2.205E0	1.3

Conclusions…

It is possible to overcome the problem of multiple formats and layouts.



set of semantic data models

(one for each measurement type)

Data Interfaces are the optimal solution to the “standardization bottleneck.”

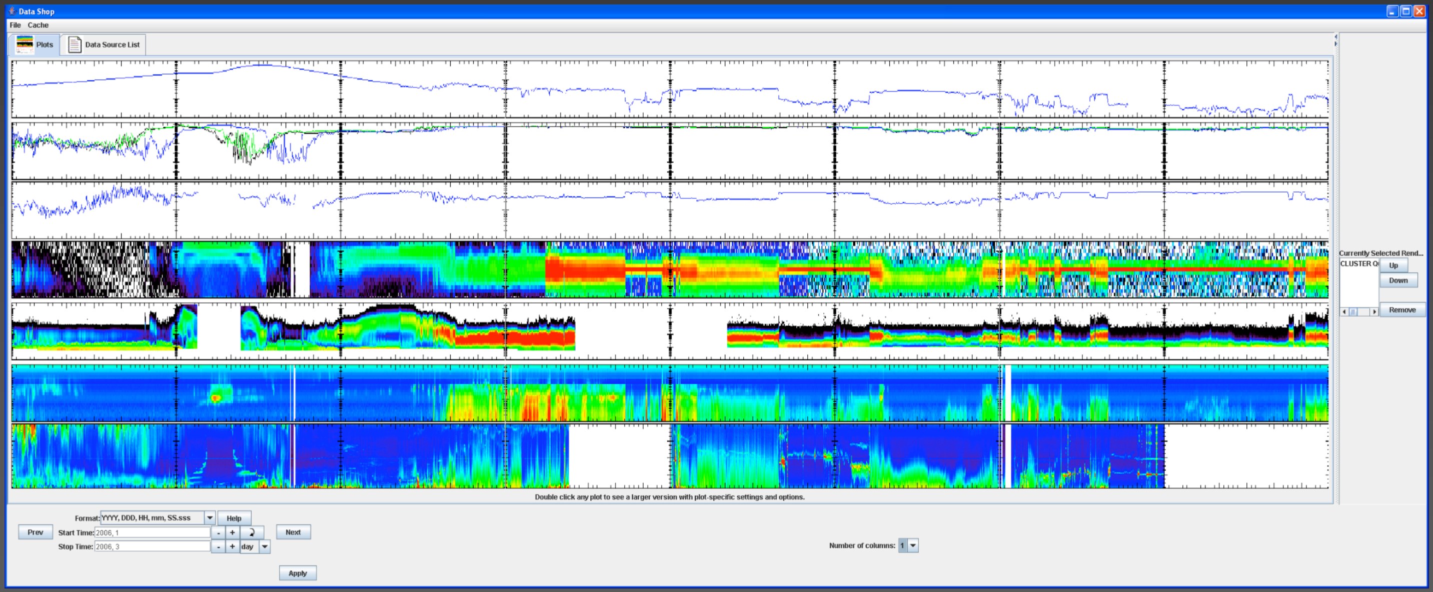
First generation systems built with this approach work well and are being expanded for use with:

VSPO – Virtual Space Physics Observatory
VHO – Virtual Heliospheric Observatory
VMO – Virtual Magnetospheric Observatory

Proof of Concept - DataShop

`http://sd-www.jhuapl.edu/datashop`

Eight Cluster 6 hour images – cropped and stitched together.



Double click any plot to see a larger version with plot-specific settings and options.

Format: YYYY, DDD, HH, mm, SS.sss

Start Time: 2006, 1

Stop Time: 2006, 3

day

Number of columns: 1

Jon Vandegriff -- JHU / APL

Benefits of Putting the Standard Inside Software

no changes to current datasets and data systems

=> all legacy formats supported; no extra work for providers

data is also transmitted in its native format (i.e., compactly)

=> the standard is NOT an XML schema, so no need to convert everything to (bulky) XML

the standard is focused on one aspect of the data – its content

=> leave messy details of data access out of the standard

data access details handled by software, not an XML schema;

=> tremendous flexibility allows all datasets to be included

you could still develop some meta-data driven interpreters

=> software need not be created for every single dataset

unique features of datasets not lost

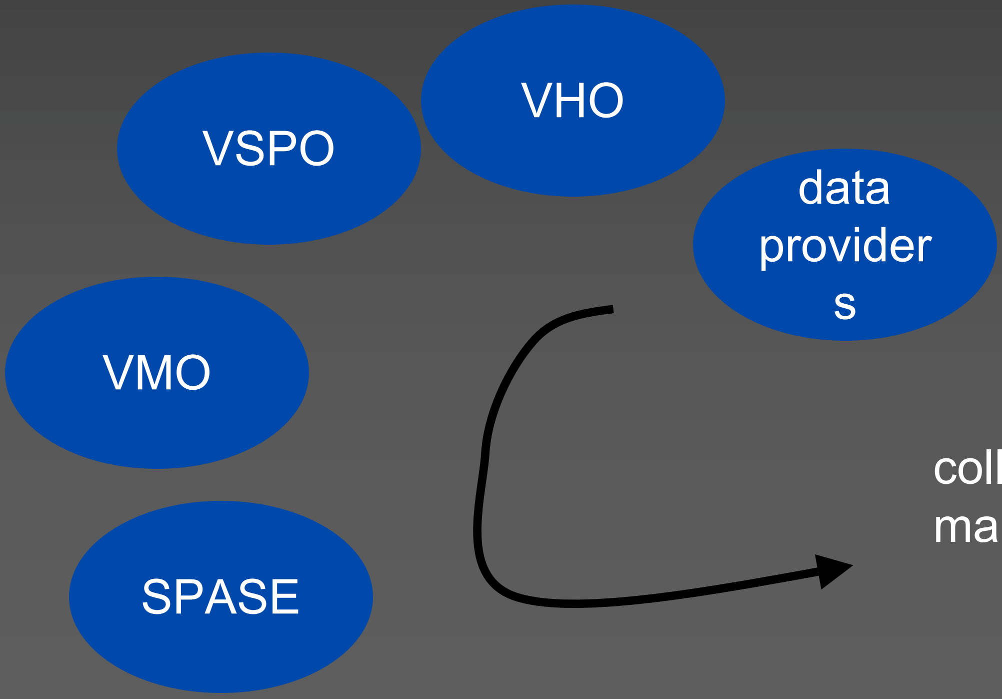
=> special features of the data are passed through the interface (but have no special semantic meaning)

How to Establish the Standards

Work with data providers and VxO's to develop a working model.

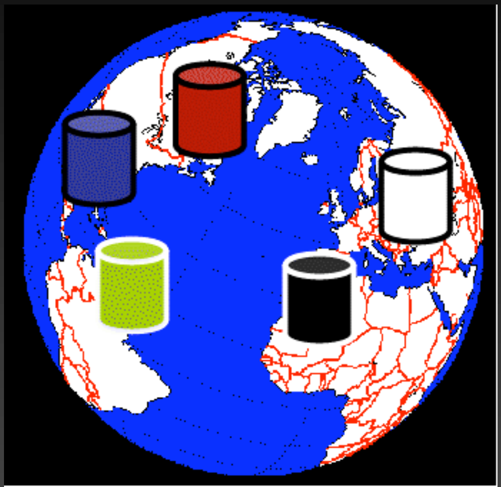
Let people see the implications of various choices in the data model, and establish a small group of people to develop the structure of the initial software library.

Acceptance of the standards will have to be market-place driven – people use and contribute to the shared library.



collect ideas from many groups

Definition: distributed heliophysical* data



-- many different types of data
-- in widely dispersed locations
-- with different access methods
-- in every possible format conceived by man
-- with many variations even within the standard formats

* mostly time-series, in-situ data

Definition: groundwork

a foundational access mechanism which enables interoperability by completely and efficiently erasing the problem of diverse data formats